EVALUATION OF A METHOD FOR SUBJECTIVE ASSESSMENT OF SPEECH QUALITY IN TELECOMMUNICATION APPLICATIONS.

Kerrie Lee, Phillip Dermody, Daniel Woo
University of Sydney

ABSTRACT: A seven point rating scale was used to obtain mean opinion scale (MOS) scores from a group of listeners who used the scale to judge the quality of speech distorted using a modulated noise reference unit (MNRU). The results showed that there were large individual differences in the use of the scale by listeners. Despite these differences individual listeners gave perceptual results which reflected the degree of MNRU distortion. An ANOVA analysis showed statistically reliable differences in the trend of the quality judgments across the MNRU conditions.

INTRODUCTION:

In the mean opinion scale (MOS) measurement method listeners hear a sample of speech and are asked to make a perceptual judgment about some aspect of the speech. Typically listeners are asked to judge the quality of the signal. The quality judgment is obtained using a scale based on a restricted range of values (typically 5 to 9 categories) which can have either a verbal description or a numeric quantity. That is, the scale may have labels such as "unsatisfactory" or "excellent" or it may have only numbers where 1 is the poorest quality and 7 is the best. The MOS test method using verbal labels can be described as an absolute category judgment as originally proposed by Wever and Zener (1928). If only numeric labels are used the method is similar to the classic rating method originally used by Galton (1883). Category rating of distorted speech material has been compared with other speech assessment methods (e.g. Mackie, Dermody and Katsch, 1987).

Studies of rating scale techniques have often focused on the question of the nature of the relationship between the underlying psychological scale (represented by the human judgments along a continuum) versus the physical scale that can be measured in the stimulus (Engen,1971). A number of practical applications have also been developed from the use of subjective rating scales. One of these is the MOS technique which has been proposed for use in a telecommunication context to provide subjective quality rating of telecommunication networking.

In order to use MOS testing to evaluate telecommunication quality there is a need to develop a set of reference materials which can be used for "calibrated" quality judgments against which the test conditions of interest can be compared. If a function can be measured indicating the quality of reference materials with known acceptability to listeners then an assignment of the quality of test conditions can be made by directly comparing the test condition results with the reference condition results.

A de facto standard of distortion in telecommunication operations is to use the quality of a single PCM process (8 bit resolution with 8 KHz sampling rate) as the base reference and to assume that distortion accumulates with increasing tandems of the PCM. Because PCMs must be implemented as devices and any one implementation may not truly represent the required additivity of several theoretical tandem PCMs an alternative method that approximates the distortion in multiple PCMs has been generally adopted. This distortion method was originally proposed by Law and Seymour (1962) and is produced by a Modulated Noise Reference Unit (MNRU) and analogue and digital implementations of this distortion have been used to assess telecommunication quality.

The present study employs the MOS method using MNRU distorted speech to investigate some underlying properties of the MOS method for determining subjective quality.

Speech Database.

The speech database consists of a recorded speech corpus comprising a set of 100 sentences spoken by 2 speakers (1 male and 1 female).in recording studio conditions via a standard telephone handset. The handset was held in a fixed position at a distance of 5cm from the speaker's mouth. The output of the telephone was recorded via a circuit that simulated a 1.6 kilometre transmission line. The sentences were analysed and equalised to some mean long term RMS level. The speech database was stored digitally for processing.

The speech materials were processed by a software MNRU algorithm that produced 5 different MNRU Q values at 33, 30 27.6. 26.5 and 24.55 dB signal to noise ratios. The reference set thereby included 1000 sentences (500 male and 500 female) spanning five distortion conditions. These materials were randomised for presentation to listeners.

MOS testing.

Subjective assessments were carried out in a sound treated room meeting audiometric testing standards. Listeners were seated at individual carrels and listened to speech through TDH-39 earphones. After each sentence presentation listeners responded on a button box showing a 7 point scale. They were instructed to indicate their judgment about the quality of the sentence that had been presented. The 7 point scale was numbered from 1 to 7 and labeled from "unsatisfactory" to "excellent". Listeners were instructed that there was no "correct" answer and that they should develop their own criteria and respond accordingly. No definition of quality was provided but it was indicated that the material should be judged relative to telephone speech.

All listeners were given 2 hours of training on the materials and method. The training data were not used in the analysis. The present data analysis is based on testing over two subsequent days with only 1 speaker presented on each day.

Listeners

MOS test results were obtained from 24 adult listeners aged between 18 and 40 years who had not participated previously in listening experiments. The listeners all had normal hearing demonstrated by standard audiometric assessment.

MOS analysis

Each listeners' MOS judgments for each sentence in each reference condition was automatically recorded by computer software which read the response buttons after each trial. These data were transferred to a statistical package for analysis. Two major issues were considered in the analysis. The first was the investigation of between listener differences in their judgments and the second was the statistical reliability of differences in the quality between the reference conditions. To discuss the first of these issues the results of the 24 listeners are presented for the male speaker.

For the investigation of statistical differences between the reference conditions an overall two way ANOVA (speaker by MNRU condition) was followed by a set of post hoc comparisons using a Scheffe analysis. The significance level of the post hoc comparisons was adjusted using a Bonferroni correction to control error effects in multiple comparisons. All reported significant results are at least at 0.05 level.

RESULTS

The results of the individual MOS scores for each reference condition are reported in Figure 1 which shows the results for the 24 listeners for the male speaker. The listeners were randomly divided into 2 groups which are graphed separately for ease of viewing the individual functions Figure 1 demonstrates two interesting aspects of the MOS results. First the MOS results show a consistent monotonic relationship between MOS scores and speech distortion conditions. As the MNRU distortion increases (1 is the least distorted and 5 is the most distorted) the MOS score consistently decreases (a score of 7 is highest perceptual quality and 1 is poorest perceptual quality). This suggests that a MOS test provides very reliable judgments even for inexperienced listeners with minimum training and even for fairly subtle differences between distortion conditions. That is, even though the difference between reference conditions is small, for example the difference between reference conditions 4 and 5 is only 1.1 dB the listeners are able to respond to this difference in a perceptual rating scale).

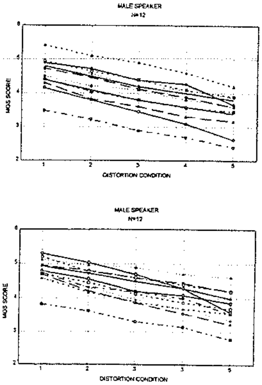

Figure 1: Mean MOS score for each test distortion for individual listeners in the male speaker condition.

The second interesting effect in Figure 1 is the limited number of MOS categories used by all listeners and the individual boundaries used to capture these compressed scales. That is, all listeners seem to compress their judgments to a limited number of choices on the scale but the points on the scale over which this compression operates is different for each listener. That is, some listeners may not use the lower part of the scale which other listeners may not use either extreme of the scale. One reason that a 7 point scale was chosen for this testing was based on our observations that even experienced listeners use a compressed scale and that a 7 point scale may provide a greater sensitivity to differences than a 5 point rating scale. We do not have any substantial data using 5 point scales for inexperienced listeners but the present results indicate that any additional compression of the category scale may lead to a very close bunching of responses in the testing and therefore reduced sensitivity.

The large individual differences in the range of MOS scores used to make quality judgments suggest that the variance around the means of MOS tests is artificially large and that a normalising procedure for the responses might be the most appropriate method to use. Alternatively, use of individual results for comparison conditions with the references might provide the most reliable estimates of category judgments for test procedures. The direct comparison of test and reference conditions would then be calculated for each individual not for the group MOS function and summed only at the end to provide an average rating of test condition.

On the other hand because there are a large number of quality estimates for each MNRU condition and because the trend of the means demonstrates a consistently decreasing function with a similar variance around the means a more traditional ANOVA analysis might also be used to demonstrate consistent differences between distortion conditions. In order to confirm this speculation an ANOVA analysis was carried out on the data for 24 listeners over the MNRU conditions with speaker gender as an analysis factor.

Figure 2 shows the average MOS results for all listeners across MNRU conditions for the male and female speakers. The ANOVA results for the means presented in Figure 2 indicate a significant effect for MNRU condition with no effect for speaker gender. The overall significance of the effect of MNRU condition indicated the appropriateness of carrying out post hoc comparisons to test differences between the individual MNRU condition means.

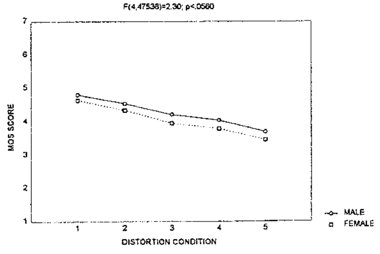

Figure 2. Mean MOS score for each reference distortion condition for male and female speaker

The post hoc comparisons produced a significant difference between each ordered set of means (that is between distortion condition 1 and distortion condition 2, distortion condition 2 and distortion condition 3 etc.). This result indicates that the trend observed in the averaged results in Figure 2 is reliable and that MNRU distortion produces a consistently decreasing function for the MOS scores.

DISCUSSION

The present data suggest that the MOS test technique provides a reasonably consistent set of results for quality judgments of MNRU distorted speech. Listeners restrict their response choices to a limited subset of the range MOS score possibilities and each listener uses different boundaries for this restricted set. However the MOS results produce statistically reliable differences between the MNRU conditions tested in the study at least for the number of responses obtained in the use of large number of listeners and test sentences.

The results of this study indicate that the MNRU conditions provide a reliable set of reference conditions. These reference conditions could have an application in defining acceptability levels for test devices based on their relationship with degree of MNRU distortion.

One outstanding issue is whether the comparison of test conditions with the MNRU reference set might be more valid if carried out for individual results (i.e. directly comparing the test results with the reference results for each individual before averaging to provide an overall result for a test condition) or whether the more traditional ANOVA method will be adequate for deciding whether a test condition is statistically different from a reference condition.

Another issue is the suitability of using the MNRU distortion references against which to compare a wide range of speech processing algorithms. While we have started to collect data about these questions we have not progressed far enough to report them in detail.

REFERENCES:

Engen, T. (1971) "Psychophysics. Scaling Methods". In Kling, J. and Riggs, L. (Eds.) *Woodworth and Schlosberg's Experimental Psychology*. London: Methuen. Pp. 47-86.

Galton, F. (1883) *Inquiries into Human Faculty and its Development*. London: Macmillan.

Law, H. and Seymour, R. (1962) "A reference distortion system using modulated noise". IEEE, November, 484-485.

Mackie (Lee), K., Dermody, P. and Katsch, R. (1987) "Assessment of evaluation measures for processed speech." Speech Communication, 6, 1-9.

Wever, E. and Zener, K. (1928) "The method of absolute judgment in psychophysics". Psychological Review, 35, 466-493.